Slowness: An Objective for Spike-Timing–Dependent Plasticity?

Henning Sprekeler*, **Christian Michaelis, Laurenz Wiskott**

Institute for Theoretical Biology, Humboldt-Universität zu Berlin, Berlin, Germany

Our nervous system can efficiently recognize objects in spite of changes in contextual variables such as perspective or lighting conditions. Several lines of research have proposed that this ability for invariant recognition is learned by exploiting the fact that object identities typically vary more slowly in time than contextual variables or noise. Here, we study the question of how this "temporal stability" or "slowness" approach can be implemented within the limits of biologically realistic spike-based learning rules. We first show that slow feature analysis, an algorithm that is based on slowness, can be implemented in linear continuous model neurons by means of a modified Hebbian learning rule. This approach provides a link to the trace rule, which is another implementation of slowness learning. Then, we show analytically that for linear Poisson neurons, slowness learning can be implemented by spike-timing–dependent plasticity (STDP) with a specific learning window. By studying the learning dynamics of STDP, we show that for functional interpretations of STDP, it is not the learning window alone that is relevant but rather the convolution of the learning window with the postsynaptic potential. We then derive STDP learning windows that implement slow feature analysis and the "trace rule." The resulting learning windows are compatible with physiological data both in shape and timescale. Moreover, our analysis shows that the learning window can be split into two functionally different components that are sensitive to reversible and irreversible aspects of the input statistics, respectively. The theory indicates that irreversible input statistics are not in favor of stable weight distributions but may generate oscillatory weight dynamics. Our analysis offers a novel interpretation for the functional role of STDP in physiological neurons.

Introduction

The ability to recognize objects in spite of possible changes in position, lighting conditions, or perspective is doubtlessly an advantage in everyday life. However, our brain usually performs this task with such astonishing ease that we are seldom aware of the complexity this recognition problem comprises. On the level of primary sensory signals (e.g., light that stimulates a single retinal receptor), even small changes in the position of the object to be recognized may lead to vastly different stimuli. Our brain thus has to somehow identify rather different stimuli as representations of the same underlying cause, i.e., it has to develop an internal representation that is invariant to irrelevant changes of the stimulus. The work presented here is motivated by the question of how such invariant representations could be established.

Because of the limited amount of information in the genome as well as the apparent flexibility of the neural development in different environments, it seems unlikely that the information needed to form invariant representations is already there at the beginning of individual development. Some information must be gathered from the sensory input experienced during interaction with the environment; it has to be learned. As this learning process is likely to be at least partially unsupervised, the brain requires a heuristics as to what stimuli should be classified as being the same.

One possible indicator for stimuli to represent the same object is temporal proximity. A scene that the eye views is very unlikely to change completely from one moment to the next. Rather, there is a good chance that an object that can be

seen now will also be present at the next instant of time. This implies that invariant representations should remain stable over time, that is, they should vary slowly. Inverting this reasoning, a sensory system that adapts to its sensory input in order to extract slowly varying aspects may succeed in learning invariant representations. This "slowness" or "temporal stability" principle is the basis of a whole class of learning algorithms [1–7]. Most applications of this approach have focused on models of the visual system, in particular on the self-organized formation of complex cell receptive fields in the primary visual cortex [8,9].

For clarity, we will focus on one of these algorithms, slow feature analysis (SFA; [10]); a close link to the so-called "trace rule" will arise naturally. The goal of SFA is the following: given a multidimensional input signal $\mathbf{x}(t)$ and a finite-dimensional function space $\mathcal{F}$, find the input–output function $g_1(\mathbf{x})$ in $\mathcal{F}$ that generates the most slowly varying output signal $y_1(t) = g_1(\mathbf{x}(t))$. It is important to note that the function $g_1(\mathbf{x})$ is required to be an instantaneous function of the input

Abbreviations: EPSP, excitatory postsynaptic potential; LTD, long-term depression; LTP, long-term potentiation; SFA, slow feature analysis; STDP, spike-timing–dependent plasticity

* To whom correspondence should be addressed. E-mail: h.sprekeler@biologie. hu-berlin.de

Author Summary

Neurons interact by exchanging information via small connection sites, so-called synapses. Interestingly, the efficiency of synapses in transmitting neuronal signals is not static, but changes dynamically depending on the signals that the associated neurons emit. As neurons receive thousands of synaptic input signals, they can thus "choose" the input signals they are interested in by adjusting their synapses accordingly. This adaptation mechanism, known as synaptic plasticity, has long been hypothesized to form the neuronal correlate of learning. It raises a difficult question: what aspects of the input signals are the neurons interested in, given that the adaptation of the synapses follows a certain mechanistic rule? We address this question for spike-timing–dependent plasticity, a type of synaptic plasticity that has raised a lot of interest in the last decade. We show that under certain assumptions regarding neuronal information transmission, spike-timing–dependent plasticity focuses on aspects of the input signals that vary slowly in time. This relates spike-timing–dependent plasticity to a class of abstract learning rules that were previously proposed as a means of learning to recognize objects in spite of contextual changes such as size or position. Based on this link, we propose a novel functional interpretation of spike-timing–dependent plasticity.

signal. Otherwise, slow output signals could be generated by low-pass filtering the input signal. As the goal of the slowness principle is to detect slowly varying features of the *input* signals, a mere low-pass filter would certainly generate slow output signals, but it would not serve the purpose.

As a measure of slowness, or rather "fastness," SFA uses the variance of the time derivative, $\langle \dot{y}(t)^2 \rangle_t$, which is the objective function to be minimized. Here, $\langle \cdot \rangle_t$ denotes temporal averaging. For mathematical convenience and to avoid the trivial constant response, $y_1(t) = \text{const}$, a zero-mean, and unit variance constraint are imposed. Furthermore, it is possible to find a second function $g_2(\mathbf{x})$ extracting $y_2(t) = g_2(\mathbf{x}(t))$ that again minimizes the given objective under the constraint of being uncorrelated with $y_1(t)$, a third one uncorrelated with both $y_1(t)$ and $y_2(t)$, and so on, thereby generating a set of slow features of the input ordered by the degree of slowness. However, in this paper, we will consider just one single output unit.

SFA has been applied to the learning of translation, rotation, and other invariances in a model of the visual system [10], and it has been shown that when applied to image sequences generated from static natural images, SFA learns functions that reproduce a wide range of features of complex cells in primary visual cortex [8]. Iteration of the same principle in a hierarchical model in combination with a sparseness objective has been used to model the self-organized formation of spatial representations resembling place cells as found in the hippocampal formation of rodents [11] (see [12] for related work).

These findings suggest that on an abstract level SFA reflects certain aspects of cortical information processing. However, SFA as a technical algorithm is biologically rather implausible. There is in particular one step in its canonical formulation that seems especially odd compared with what neurons are normally thought to do. In this step the eigenvector that corresponds to the smallest eigenvalue of the covariance matrix of the time derivative of some

multidimensional signal is extracted. The aim of this paper is to show how this kind of computation can be realized in a spiking model neuron.

In the following, we will first consider a continuous model neuron and demonstrate that a modified Hebbian learning rule enables the neuron to learn the slowest (in the sense of SFA) linear combination of its inputs. Apart from providing the basis for the analysis of the spiking model, this section reveals a mathematical link between SFA and the trace learning rule, another implementation of the slowness principle. We then examine if these findings also hold for a spiking model neuron, and find that for a linear Poisson neuron, spike-timing–dependent plasticity (STDP) can be interpreted as an implementation of the slowness principle.

Results

Continuous Model Neuron

Linear model neuron and basic assumptions. First, consider a linear continuous model neuron with an input–output function given by

$$a^{\text{out}}(t) = \sum_{i=1}^{n} w_i a_i^{\text{in}}(t), \tag{1}$$

with $a_i^{in}(t)$ indicating the input signals, w_i the weights, and a^{out} the output signal. For mathematical convenience, let $a_i^{in}(t)$ and $a^{\text{out}}(t)$ be defined on the interval $t \in [-\infty, \infty]$ but differ from zero only on $[0,T]$, which could be the lifetime of the system. We assume that the input is approximately whitened on any sufficiently large interval $[t_a, t_b] \subset [0,T]$ (i.e., each input signal has approximately zero mean and unit variance and is uncorrelated with other input signals):

$$\int_{t_a}^{t_b} a_i^{\text{in}}(t)\, \mathrm{d}t \approx 0 \quad \text{(zero mean)}, \tag{2}$$

$$\frac{1}{T_{ab}} \int_{t_a}^{t_b} a_i^{\text{in}}(t)^2\, \mathrm{d}t \approx 1 \quad \text{(unit variance)}, \tag{3}$$

$$\int_{t_a}^{t_b} a_i^{\text{in}}(t)\, a_{j \neq i}^{\text{in}}(t)\, \mathrm{d}t \approx 0 \quad \text{(decorrelation)}. \tag{4}$$

This can be achieved by a normalization and decorrelation step of the units projecting to the considered unit. Furthermore, we assume that the output is normalized to unit variance, which for whitened input means that the weight vector is normalized to length 1. In an online learning rule, this could be implemented by either an activity-dependent or a weight-dependent normalization term. Thus, for the output signal we have:

$$\int_{t_a}^{t_b} a^{\text{out}}(t)\, \mathrm{d}t \overset{(1,2)}{\approx} 0, \quad \text{(zero mean)} \tag{5}$$

$$\frac{1}{T_{ab}} \int\limits_{t_a}^{t_b} a^{\text{out}}(t)^2 \, dt \stackrel{(1,3)}{\approx} \sum_{i=1}^{n} w_i^2 := 1. \quad \text{(unit variance)} \quad (6)$$

In the following, we will often consider filtered signals. Therefore, we introduce abbreviations for the convolution $f \circ g$ and the cross-correlation $f * g$ of two functions $f(t)$ and $g(t)$:

$$\text{Convolution: } [f \circ g](t) := \int\limits_{-\infty}^{\infty} f(\tau)g(t-\tau) \, d\tau, \quad (7)$$

$$\text{Cross-correlation: } [f * g](t) := \int\limits_{-\infty}^{\infty} f(\tau)g(t+\tau) \, d\tau. \quad (8)$$

For convenience, we will often use windowed signals, indicated by a hat

$$\hat{s}(t) = \begin{cases} s(t) & \text{for} \quad t \in [t_a, t_b] \\ 0 & \text{otherwise} \end{cases}, \quad (9)$$

which allows us to replace the integration of a signal $s(t)$ over $[t_a, t_b]$ by an integration of $\hat{s}(t)$ over $[-\infty, \infty]$. We assume that the interval $[t_a, t_b]$ is long compared to the width of the filters. In this case, effects from the integration boundaries are negligible, and we have

$$\int\limits_{t_a}^{t_b} [f \circ s](t) \; h(t) \, dt \approx \int\limits_{-\infty}^{\infty} [f \circ \hat{s}](t) h(t) \, dt. \quad (10)$$

Similar considerations hold for the cross-correlation (Equation 8).

Since convolution and cross-correlation are conveniently treated in Fourier space, we repeat the definition of the Fourier transform $\mathcal{F}_s(\nu)$ and the power spectrum $P_s(\nu)$ of a signal $s(t)$.

$$\text{Fourier transform: } s(t) =: \int\limits_{-\infty}^{\infty} \mathcal{F}_s(\nu) \, e^{2\pi i \nu t} \, d\nu, \quad (11)$$

$$\text{Power spectrum: } \mathcal{P}_s(\nu) := \mathcal{F}_s(\nu)\bar{\mathcal{F}}_s(\nu). \quad (12)$$

Throughout the paper, we make the assumption that input signals (and hence also the output signals) do not have significant power above some reasonable frequency ν_{max}.

Reformulation of the slowness objective. SFA is based on the minimization of the second moment of the time derivative, $\int \dot{a}^{\text{out}}(t)^2 \, dt$. Even though there are neurons with transient responses to changes in the input, we believe it would be more plausible if we could derive an SFA-learning rule that does not depend on the time derivative, because it might be difficult to extract, especially for spiking neurons. It is indeed possible to replace the time derivative by a low-pass filtering as follows:

$$\text{minimize} \int\limits_{-\infty}^{\infty} \dot{a}^{\text{out}}(t)^2 \, dt \quad (13)$$

$$= \int\limits_{-\infty}^{\infty} \mathcal{P}_{\dot{a}^{\text{out}}}(\nu) \, d\nu \text{ (because of Parseval's theorem)} \quad (14)$$

$$= 4\pi^2 \int\limits_{-\infty}^{\infty} \nu^2 \mathcal{P}_{a^{\text{out}}}(\nu) \, d\nu \text{ (since } \mathcal{F}_{\dot{s}}(\nu) = 2\pi i \nu \mathcal{F}_s(\nu)) \quad (15)$$

$$\Leftrightarrow \text{maximize} \int\limits_{-\infty}^{\infty} -\nu^2 \mathcal{P}_{a^{\text{out}}}(\nu) \, d\nu \quad (16)$$

$$\Leftrightarrow \text{maximize} \int\limits_{-\infty}^{\infty} (\nu_{\max}^2 - \nu^2)\mathcal{P}_{a^{\text{out}}}(\nu) \, d\nu$$

$$\left(\text{since} \int\limits_{-\infty}^{\infty} \mathcal{P}_{a^{\text{out}}}(\nu) \, d\nu = \int\limits_{-\infty}^{\infty} a^{\text{out}}(t)^2 \, dt \stackrel{(6)}{\approx} \text{const} \right) \quad (17)$$

$$= \int\limits_{-\infty}^{\infty} \max(0, (\nu_{\max}^2 - \nu^2))\mathcal{P}_{a^{\text{out}}}(\nu) \, d\nu \quad (18)$$

$$\text{(since } \mathcal{P}_{a^{\text{out}}}(\nu) = 0 \text{ for } |\nu| > \nu_{\max} \text{ by assumption)}$$

$$= \int\limits_{-\infty}^{\infty} \mathcal{P}_{f_{\text{SFA}}}(\nu)\mathcal{P}_{a^{\text{out}}}(\nu) \, d\nu \quad (19)$$

$$\left(\text{with } f_{\text{SFA}}(t) \text{ defined such that } \mathcal{P}_{f_{\text{SFA}}}(\nu) = \max(0, (\nu_{\max}^2 - \nu^2)) \right) \quad (20)$$

$$= \int\limits_{-\infty}^{\infty} [f_{\text{SFA}} \circ a^{\text{out}}](t)^2 \, dt. \quad (21)$$

Thus, SFA can be achieved either by minimizing the variance of the time derivative of the output signal or by maximizing the variance of the appropriately filtered output signal. Figure 1 provides an intuition for this alternative. The filter f_{SFA} is obviously a low-pass filter, as one would expect, with a $\nu_{\max}^2 - \nu^2$ power spectrum below the limiting frequency $\nu_{\max}$. Because the phases are not determined, further assumptions are required to fully determine an SFA filter. However, we will proceed without defining a concrete filter, since it is not required for the considerations below.

Hebbian learning on filtered signals. It is known that standard Hebbian learning under the constraint of a unit weight vector applied to a linear unit maximizes the variance of the output signal. We have seen in the previous section that SFA can be reformulated as a maximization problem for the variance of the low-pass filtered output signal. To achieve this, we simply apply Hebbian learning to the filtered input and output signals, instead of to the original signals.

Consider a hypothetical unit that receives low-pass filtered inputs and, therefore, because of the linearity of the unit and the filtering, generates a low-pass filtered output
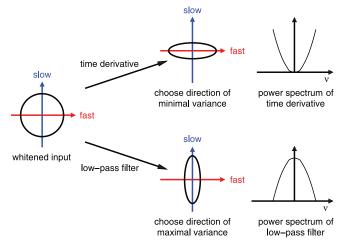
Figure 1. Choosing Slow Directions of the Input
Finding the direction of least variance in the time derivative of the input (which is part of the SFA algorithm) can be replaced by finding the direction of maximum variance in an appropriately low-pass filtered version of the input signal.
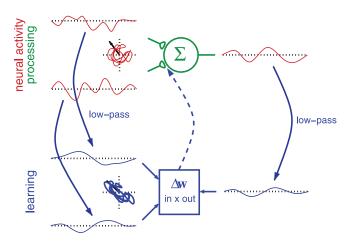doi:10.1371/journal.pcbi.0030112.g001

Figure 2. Filtered Hebbian Learning Rule
Input and output signals are filtered (downward arrows). The weight change is the result of applying the Hebbian learning rule on the filtered signals (square box and upward arrow). Thereby, the variance of the filtered version of the output is maximized without actually filtering the output during processing.
doi:10.1371/journal.pcbi.0030112.g002

$$[f_{\text{SFA}} \circ a^{\text{out}}](t) \stackrel{(1)}{=} \left[f_{\text{SFA}} \circ \sum_{i=1}^{n} w_i a_i^{\text{in}} \right](t) = \sum_{i=1}^{n} w_i [f_{\text{SFA}} \circ a_i^{\text{in}}](t), \tag{22}$$

where f_{SFA} is the kernel of the linear filter applied. It is obvious that a *filtered Hebbian learning rule*

$$\dot{w}_i = \gamma \left[f^{\text{in}} \circ a_i^{\text{in}} \right](t) [f^{\text{out}} \circ a^{\text{out}}](t) \tag{23}$$

with $f^{\text{in}} := f^{\text{out}} := f_{SFA}$ maximizes the objective in Equation 21.

Remember that the input is white (i.e., the a_i^{in} are uncorrelated and have unit variance), and the weight vector is normalized to norm one by some additional normalization rule, so that we know that the output signal a^{out} has the same variance no matter what the direction of the weight vector is. Thus, the filtered Hebbian plasticity rule (together with the normalization rule not specified here) optimizes slowness (Equation 13) under the constraint of unit variance (Equation 6). Figure 2 illustrates this learning scheme. It also underlines the necessity for a clear distinction between processing and learning. Although the slowness principle does not allow low-pass filtering as a means of generating slow signals during processing, the learning rule may well make use of low-pass filtered signals to detect slowly varying features in the input signal. This distinction will become particularly important for the Poisson model neuron below, as it incorporates an excitatory postsynaptic potential (EPSP) that acts as a low-pass filter during processing. An implementation of the slowness principle in such a system must avoid the system exploiting the EPSP as a means of generating slow signals.

Alternative filtering procedures. If learning is slow, the total weight change over a time interval $[t_a, t_b]$ in a synapse can be written as

$$\Delta w_i := \int_{t_a}^{t_b} \dot{w}_i(t) \, dt \tag{24}$$

$$\stackrel{(23)}{=} \gamma \int_{t_a}^{t_b} \left[f^{\text{in}} \circ a_i^{\text{in}} \right](t) [f^{\text{out}} \circ a^{\text{out}}](t) \, dt \tag{25}$$

$$\stackrel{(10)}{\approx} \gamma \int_{-\infty}^{\infty} \left[f^{\text{in}} \circ \hat{a}_i^{\text{in}} \right](t) [f^{\text{out}} \circ \hat{a}^{\text{out}}](t) \, dt \tag{26}$$

$$= \gamma \int_{-\infty}^{\infty} [[f^{\text{out}} * f^{\text{in}}] \circ \hat{a}_i^{\text{in}}](t) \, \hat{a}^{\text{out}}(t) \, dt \tag{27}$$

$$= \gamma \int_{-\infty}^{\infty} \hat{a}_i^{\text{in}}(t) [[f^{\text{in}} * f^{\text{out}}] \circ \hat{a}^{\text{out}}](t) dt \tag{28}$$

$$= \gamma \int_{-\infty}^{\infty} [f^{\text{in}} * f^{\text{out}}](t) [\hat{a}^{\text{out}} * \hat{a}_i^{\text{in}}](t) \, dt. \tag{29}$$

Thus, one can either convolve input and output signal with filters f^{in} and f^{out}, respectively, the input signal with $f^{\text{out}} * f^{\text{in}}$, or the output signal with $f^{\text{in}} * f^{\text{out}}$. Note that $[f^{\text{in}} * f^{\text{out}}](t) = [f^{\text{out}} * f^{\text{in}}](-t)$. One can actually use any pair of filters f^{in} and f^{out} as long as $f^{\text{in}} * f^{\text{out}}$ fulfills the condition

$$\mathcal{F}_{f^{\text{in}} * f^{\text{out}}}(\nu) = \mathcal{P}_{f_{\text{SFA}}}(\nu). \tag{30}$$

Relation to other learning rules. Hebbian learning on low-pass filtered signals is the basis of several other models for unsupervised learning of invariances [1,4,6]. These models essentially subject the output signal to an exponential temporal filter $f(t) = \theta(t)\exp(-\gamma t)$ and then use Hebbian learning to associate it with the input signal. Here, $\theta(t)$ denotes the Heaviside step function, which is 0 for $t < 0$ and 1 for $t \geq 0$. This learning rule has been named the "trace rule." The considerations in the last section provide a link between

this approach and ours. We simply have to replace f^{in} with a δ-function and f^{out} with $f(t)$. Equation 29 then takes the form

$$\Delta w_i = \gamma \sum_j \left[\int_{-\infty}^{\infty} f(t)[\hat{a}_j^{\text{in}} * \hat{a}_i^{\text{in}}](t)\, \mathrm{d}t \right] w_j, \tag{31}$$

since the output signal $a^{\text{out}} = \sum_j w_j a_j^{\text{in}}$ is a linear function of the input (see Equation 1). In the previously mentioned applications of the trace rule, the statistics of the input signals were always reversible, so we will assume that all correlation functions $[\hat{a}_i^{\text{in}} * \hat{a}_j^{\text{in}}](t)$ are symmetric in time. This implies that only the symmetric component of $f(t)$ is relevant for learning:

$$f^{\text{sym}}(t) := \frac{1}{2}(f(t) + f(-t)) = \frac{\gamma}{2}\exp(-\gamma|t|). \tag{32}$$

It is easy to show that the learning rule in Equation 31 can be interpreted as a gradient ascent on the following objective function:

$$\Psi = \int_{-\infty}^{\infty} f^{\text{sym}}(t)[\hat{a}^{\text{out}} * \hat{a}^{\text{out}}](t)\, \mathrm{d}t \tag{33}$$

$$= \int_{-\infty}^{\infty} \mathcal{F}_{f^{\text{sym}}}(\nu)\, \mathcal{P}_{\hat{a}^{\text{out}}}(\nu)\, \mathrm{d}\nu. \tag{34}$$

By comparison with Equation 19, it becomes clear that the trace rule implements a very similar objective as our model. The only difference is that the power spectrum in Equation 20 is replaced by the Fourier transform of the filter f^{sym}. Note that in order to be able to interpret Ψ as an objective function, it should be real-valued. The replacement of f with f^{sym} ensures that $\mathcal{F}_{f^{\text{sym}}}$ is real-valued and symmetric, so Ψ is real-valued as well. The Fourier transform of f^{sym} is given by

$$\mathcal{F}_{f^{\text{sym}}}(\nu) = \frac{\gamma}{\gamma^2 + (2\pi\nu)^2}. \tag{35}$$

This shows that the only difference between the trace rule and our model lies in the choice of the power spectrum for the low-pass filter. While we are using a parabolic power spectrum with a cutoff (Equation 20), the trace rule uses a power spectrum with the shape of a Cauchy function (Equation 35).

From this perspective, one can interpret SFA as a quadratic approximation of the trace rule. To what extent this approximation is valid depends on the power spectra of the input signals. If most of the input power is concentrated at low frequencies, where the power spectrum resembles a parabola, the learning rules can be expected to learn very similar weight vectors. In fact, any Hebbian learning rule that leads to an objective function of the shape of Equation 19 with a low-pass filtering spectrum in the place of $\mathcal{P}_{f_{\text{SFA}}}$ essentially implements the slowness principle, as among signals with the same variance, it will favor slower ones.

Spiking Model Neuron

Real neurons do not transmit information via a continuous stream of analog values like the model neuron considered in the previous section, but rather emit action potentials that carry information by means of their rate and probably also by their exact timing, a fact we will not consider here. How can the model developed so far be mapped onto this scenario?

The linear Poisson neuron. Again, we restrict our analysis to a simple case by modeling the spike-train signals by inhomogeneous Poisson processes. Note that at this point, we restrict our analysis to a rate code, thus neglecting possible coding paradigms that rely on precise timing of spikes.

To generate the input spike trains, we first add sufficiently large constants c_i^{in} to the continuous and zero-mean signals $a_i^{\text{in}}(t)$ to turn them into strictly positive signals that can be interpreted as rates

$$r_i^{\text{in}}(t) := c_i^{\text{in}} + a_i^{\text{in}}(t). \tag{36}$$

The constants c_i^{in} represent mean firing rates, which are modulated by the input signals a_i^{in}. From the input rates $r_i^{\text{in}}(t)$, we then derive inhomogeneous Poisson spike trains $S_i^{\text{in}}(t)$ drawn from ensembles E_i^{in} such that

$$\langle S_i^{\text{in}}(t) \rangle_{E_i^{\text{in}}} = r_i^{\text{in}}(t), \tag{37}$$

where $\langle \cdot \rangle_{E_i^{\text{in}}}$ denotes the average over the ensemble E_i^{in}.

The output rate is modeled as a weighted sum over the input spike trains convolved with an EPSP $\varepsilon(t)$ plus a baseline firing rate r_0, which ensures that the output firing rate remains positive. This is necessary as we allow inhibitory synapses (i.e., negative weights).

$$m(t) := r_0 + \sum_{i=1}^{n} w_i \left[\varepsilon \circ S_i^{\text{in}} \right](t) \tag{38}$$

Note that in this scheme, the EPSP reflects the change in the postsynaptic firing probability due to a presynaptic spike rather than a change in the membrane potential. Ideally, it includes all delay effects in neuronal transmission.

The output of this spiking neuron is yet another inhomogeneous Poisson spike train $S^{\text{out}}(t)$ drawn from an ensemble E^{out}, given a realization of the input spike trains S_i^{in} such that

$$\langle S^{\text{out}}(t) \rangle_{E^{\text{out}}|\{S_i^{\text{in}}\}} = m(t). \tag{39}$$

It should be noted that not only is the output spike train $S^{\text{out}}(t)$ stochastic in this model, but also the underlying output rate $m(t)$, which is a function of the stochastic variables $S_i(t)$ and generally differs for each realization of the input. This is the reason why the input and output spike trains are not statistically independent. However, due to the linearity of the model neuron, the output rate is still simply

$$r^{\text{out}}(t) = \langle S^{\text{out}}(t) \rangle_{E_i^{\text{in}}, E^{\text{out}}} \tag{40}$$

$$\overset{(39,38,37)}{=} r_0 + \sum_{i=1}^{n} w_i \left[\varepsilon \circ r_i^{\text{in}} \right](t) \tag{41}$$

$$\overset{(36)}{=} \underbrace{r_0 + \sum_{i=1}^{n} w_i c_i^{\text{in}} \int_{-\infty}^{\infty} \varepsilon(t)\, \mathrm{d}t}_{=:c^{\text{out}}} + \sum_{i=1}^{n} w_i \left[\varepsilon \circ a_i^{\text{in}} \right](t) \tag{42}$$

$$= c^{\text{out}} + \left[\varepsilon \circ \sum_{i=1}^{n} w_i a_i^{\text{in}} \right](t) \tag{43}$$

$$\overset{(1)}{=} c^{\mathrm{out}} + [\varepsilon \circ a^{\mathrm{out}}](t), \qquad (44)$$

and the joint firing rate is

$$r_i^{\mathrm{in,out}}(t,t') := \langle S_i^{\mathrm{in}}(t)\,S^{\mathrm{out}}(t') \rangle_{E_i^{\mathrm{in}},E^{\mathrm{out}}} \qquad (45)$$

$$= r_i^{\mathrm{in}}(t) r^{\mathrm{out}}(t') + w_i \varepsilon(t'-t) r_i^{\mathrm{in}}(t) \quad (\text{see } [13]). \qquad (46)$$

The first term would result also from a rate model, while the second term captures the statistical dependencies between input and output spike trains mediated by the synaptic weights w_i and the EPSP $\varepsilon(t)$.

STDP can perform SFA. In this section, we will demonstrate that in an ensemble-averaged sense it is possible to generate the same weight distribution as in the continuous model by means of an STDP rule with a specific learning window.

Synaptic plasticity that depends on the temporal order of pre- and postsynaptic spikes has been found in a number of neuronal systems [14–18], and has raised a lot of interest among modelers [19,20] (for a review, see [21]). Typically, synapses undergo long-term potentiation (LTP) if a presynaptic spike precedes a postsynaptic spike within a timescale of tens of milliseconds and long-term depression (LTD) for the opposite temporal order. Assuming that the change in synaptic efficacy occurs on a slower timescale than the typical interspike interval, the STDP weight dynamics can be modeled as

$$\Delta w_i = \gamma \sum_{\alpha}^{m_i^{\mathrm{in}}} \sum_{\beta}^{m^{\mathrm{out}}} W(t_{i\alpha}^{\mathrm{in}} - t_{\beta}^{\mathrm{out}}). \qquad (47)$$

Here, $t_{i\alpha}^{\mathrm{in}}$ denotes the spike times of the presynaptic spikes at synapse i and t_{β}^{out} denotes the postsynaptic spike times. $W(t)$ is the learning window that determines if and to what extent the synapse is potentiated or depressed by a single spike pair. The convention is such that negative arguments t in $W(t)$ correspond to the situation where the presynaptic spike precedes the postsynaptic spike. m_i^{in} and m^{out} are the numbers of pre- and postsynaptic spikes occurring in the time interval $[t_a, t_b]$ under consideration. γ is a small positive learning rate. Note that due to the presence of this learning rate, the absolute scale of the learning window W is not important for our analysis.

We circumvent the well-known stability problem of STDP by applying an explicit weight normalization ($\mathbf{w}^{\mathrm{new}} = (\mathbf{w}^{\mathrm{old}} + \Delta\mathbf{w})/\|\mathbf{w}^{\mathrm{old}} + \Delta\mathbf{w}\|$) instead of weight-dependent learning rates as used elsewhere [22–24]. Such a normalization procedure could be implemented by means of a homeostatic mechanism targeting the output firing rate (e.g., by synaptic scaling; for reviews, see [25,26]).

Modeling the spike trains as sums of delta pulses (i.e., $S^{\mathrm{in/out}}(t) = \sum_j \delta(t - t_j^{\mathrm{in/out}})$), the learning rule in Equation 47 can be rewritten as

$$\Delta w_i = \gamma \int_{t_a}^{t_b} \int_{t_a}^{t_b} W(t - t') S_i^{\mathrm{in}}(t) S^{\mathrm{out}}(t')\, \mathrm{d}t\, \mathrm{d}t' \qquad (48)$$

$$\approx \gamma \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W(t - t') \hat{S}_i^{\mathrm{in}}(t) \hat{S}^{\mathrm{out}}(t')\, \mathrm{d}t\, \mathrm{d}t'. \qquad (49)$$

Taking the ensemble average allows us to retrieve the rates that underlie the spike trains and thus the signals $\hat{a}_i^{\mathrm{in}}$ and $\hat{a}^{\mathrm{out}}$ of the continuous model:

$$\langle \Delta w_i \rangle_{E^{\mathrm{in}},E^{\mathrm{out}}} \overset{(49)}{\approx} \gamma \int_{-\infty}^{-\infty} \int_{-\infty}^{-\infty} W(t - t') \,\langle \hat{S}_i^{\mathrm{in}}(t) \hat{S}^{\mathrm{out}}(t') \rangle_{E^{\mathrm{in}},E^{\mathrm{out}}}\, \mathrm{d}t\, \mathrm{d}t' \qquad (50)$$

$$\overset{(46)}{=} \gamma \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W(t - t') \big(\hat{r}_i^{\mathrm{in}}(t)\ \hat{r}^{\mathrm{out}}(t') + w_i \varepsilon(t'-t) \hat{r}_i^{\mathrm{in}}(t)\big)\, \mathrm{d}t\, \mathrm{d}t' \qquad (51)$$

$$\overset{(36,44)}{=} \gamma \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W(t - t') [c_i^{\mathrm{in}} + \hat{a}_i^{\mathrm{in}}](t) [c^{\mathrm{out}} + \varepsilon \circ \hat{a}^{\mathrm{out}}]\ (t')\mathrm{d}t\, \mathrm{d}t'$$
$$+ \gamma \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} W(t - t') w_i \varepsilon(t'-t) [c_i^{\mathrm{in}} + \hat{a}_i^{\mathrm{in}}]\ (t')\, \mathrm{d}t\, \mathrm{d}t'. \qquad (52)$$

Expanding the products in Equation 52 gives rise to a number of terms, among which only one depends on both the input and the output signal $\hat{a}_i^{\mathrm{in}}$ and $\hat{a}^{\mathrm{out}}$. Because each input signal has a vanishing mean, terms containing just one input signal lead to negligible contributions. The remaining terms depend only on the mean firing rates c_i^{in} and c^{out}:

$$\langle \Delta w_i \rangle_{E^{\mathrm{in}},E^{\mathrm{out}}} \overset{(52)}{\approx} \gamma \int_{-\infty}^{\infty} W(t - t') \hat{a}_i^{\mathrm{in}}(t) [\varepsilon \circ \hat{a}^{\mathrm{out}}](t')\, \mathrm{d}t\, \mathrm{d}t'$$
$$+ \gamma w_i c_i^{\mathrm{in}} T_{ab} \int_{-\infty}^{\infty} W(t)\, \varepsilon(-t)\, \mathrm{d}t \qquad (53)$$
$$+ \gamma c^{\mathrm{out}} c_i^{\mathrm{in}} T_{ab} \int_{-\infty}^{\infty} W(t)\, \mathrm{d}t.$$

A generalized version of Equation 53 that incorporates non-Hebbian plasticity (i.e., terms that depend on the pre/postsynaptic signals only) has been derived and discussed by Kempter et al. [27]. Regarding the effects of the input signals on learning, the decisive term is the first one. The other two are rather unspecific in that they do not depend on the properties of the input and output signals $\hat{a}_i^{\mathrm{in}}$ and $\hat{a}^{\mathrm{out}}$.

The second term alone would generate a competition between the weights: synapses that experience a higher mean input firing rate c_i^{in} grow more rapidly than those with smaller input firing rates. If we assume that the input neurons fire with the same mean firing rate, all weights grow with the same rate, so the direction of the weight vector remains unchanged. Thus, due to the explicit weight normalization, this term has no effect on the weight dynamics and can be neglected.

If the integral over the learning window is positive, the

third term in Equation 53 favors a weight vector that is proportional to the vector of the mean firing rates of the input neurons. It thus stabilizes the homogeneous weight distribution and opposes the effect of the first term, which captures correlations in the input signals. Note that this is only true if the integral over the learning window is positive; otherwise, this term introduces a competition between the weights [24,27]. One possible interpretation is that the neuron has a "default state" in which all synapses are equally strong and that correlations in the input need to surpass a certain threshold in order to be imprinted in the synaptic connections. Interestingly, this threshold is determined by the integral over the learning window, which implies that neurons that balance LTP and LTD should be more sensitive to input correlations.

An alternative possibility is that the neuron possesses a mechanism of canceling the effects of this term. From a computational perspective this would be sensible, as the mean firing rates c_i^{in} and c^{out} do not carry information about the input, neither in rate nor in a timing code. If we conceive neurons as information encoders aiming at adapting to the structure of their input, this term is thus more hindrance than help. Assuming that the neuron compensates for this term, the dynamics of the synaptic weights are governed exclusively by the correlations in the input signals as reflected by the first term. In the following, we will restrict our considerations to this term and omit the others.

Rearranging the temporal integrations, we can rewrite Equation 53 for the weight updates as

$$\langle \Delta w_i \rangle_{E^{in}, E^{out}} \overset{(53)}{\approx} \gamma \int_{-\infty}^{\infty} [W \circ \varepsilon](t)[\hat{a}^{out} * \hat{a}_i^{in}](t)\, dt. \quad (54)$$

The first conclusion we can draw from this reformulation is that for the dynamics of the learning process the convolution of the learning window with the EPSP and not the learning window alone is relevant. As discussed below, this might have important consequences for functional interpretations of the shape of the learning window.

Second, by comparison with Equation 29, it is obvious that in order to learn the same weight distribution as in the continuous model, the learning window has to fulfill the condition that

$$[W \circ \varepsilon](t) = [f^{in} * f^{out}](t) =: W_0(t) \quad (55)$$

$$\Leftrightarrow \mathcal{F}_{W \circ \varepsilon}(\nu) = \mathcal{F}_W(\nu)\mathcal{F}_\varepsilon(\nu) = \mathcal{F}_{f^{in}*f^{out}}(\nu) = \mathcal{P}_{f_{SFA}}(\nu) = \mathcal{F}_{W_0}(\nu). \quad (56)$$

Here, W_0 is the convolution of W with ε and is equal to the learning window in the limit of an infinitely short, δ-shaped EPSP. As the power spectrum $\mathcal{P}_{f_{SFA}}(\nu)$ is of course real, W_0 is symmetric in time. Note that the width of W_0 scales inversely with the width of the power spectrum $\mathcal{P}_{f_{SFA}}(\nu)$, which in turn is proportional to ν_{max}. Once the power spectrum $\mathcal{P}_{f_{SFA}}(\nu)$ and the EPSP is given, Equation 56 uniquely determines the learning window W. Because it is W_0 rather than W that determines the learning dynamics, we will refer to W_0 as the "effective learning window."

Learning windows. According to the last section, we require special learning windows to learn the slow directions

in the input. This of course raises the question of which window shapes are favorable, and in particular if these are in agreement with physiological findings.

Given the shape of the EPSP and the power spectrum $\mathcal{P}_{f_{SFA}}$, the learning window is uniquely determined by Equation 56. Remember that the only parameter in the power spectrum $\mathcal{P}_{f_{SFA}}$ is the frequency ν_{max}, above which the power spectrum of the input data was assumed to vanish. For simplicity, we model the EPSP as a single exponential with a time constant τ:

$$\varepsilon(t) = \theta(t)\, e^{-\frac{t}{\tau}}. \quad (57)$$

For this particular EPSP shape, the learning window can be calculated analytically by inverting the Fourier transform in Equation 56. The result can be written as

$$W(t) = \left[\frac{d}{dt} + \frac{1}{\tau}\right] W_0(t). \quad (58)$$

W_0 is symmetric, so its derivative is antisymmetric. Thus, the learning window is a linear combination of a symmetric and an antisymmetric component. As the width of W_0 scales with the inverse of ν_{max}, its temporal derivative scales with ν_{max}. Accordingly, the symmetry of the learning window is governed by an interplay of the duration τ of the EPSP and the maximal input frequency ν_{max}. For $\tau \ll 1 / \nu_{max}$ the learning window is dominated by W_0 and thus symmetric, whereas for $\tau \gg 1 / \nu_{max}$, the temporal derivative of W_0 is dominant, so the learning window is antisymmetric.

We have assumed that the input signals have negligible power above the maximal input frequency ν_{max}. Thus, the temporal structure of the input signals can only provide a lower bound for ν_{max}. On the other hand, exceedingly high values for ν_{max} lead to very narrow learning windows, thereby sharpening the coincidence detection and reducing the speed of learning. Moreover, it may be metabolically costly to implement physiological processes that are faster than necessary. Thus, it appears sensible to choose ν_{max} such that $1 / \nu_{max}$ reflects the fastest timescale in the input signals. Accordingly, the symmetry of the learning window is governed by the relation between the length of the EPSP and the fastest timescale in the input data. If the EPSP is short enough to resolve the fastest input components, the learning window is symmetric. If the EPSP is too long to fully resolve the temporal structure of the input (i.e., it acts as a low-pass filter), the learning window will tend to be antisymmetric.

We choose a value of $\nu_{max} = 1 / (40\ ms)$. The argument for this choice is that within a rate code, the cells that project to the neuron under consideration can hardly convey signals that vary on a faster timescale than the duration of their EPSP. It is thus reasonable to choose the time constant of the EPSP and the inverse of the cutoff frequency to have the same order of magnitude. Typical durations of cortical EPSPs are of the order of tens of milliseconds (see [28] for further references and a critical discussion), so 40 ms seems a reasonable value.

Figure 3 illustrates the connection between $\mathcal{P}_{f_{SFA}}$, W_0, the learning window, and the EPSP. It also shows the learning windows for three different durations of the EPSP, while keeping $\nu_{max} = 1 / (40\ ms)$. The oscillatory and slowly decaying tails of $W(t)$ are due to the sharp cutoff of the power spectrum $\mathcal{P}_{f_{SFA}}$ at $|\nu| = \nu_{max}$ and become less pronounced if $\mathcal{P}_{f_{SFA}}$ is smoothened out.
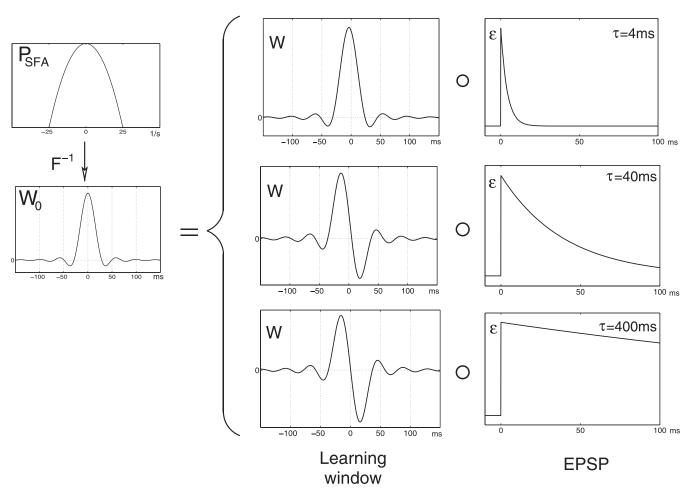
Figure 3. Relation between the EPSP and the Learning Window

The power spectrum $\mathcal{P}_{f_{SFA}}$ is the Fourier transform of the effective learning window W_0, which in turn is the convolution of the learning window W and the EPSP ε. The figure shows the learning windows required for SFA for three different EPSP durations ($\tau = 4$, 40, 400 ms). The maximal input frequency ν_{max} was 1 / (40 ms) in all plots.

doi:10.1371/journal.pcbi.0030112.g003

As negative time arguments in $W(t)$ correspond to the case in which the presynaptic spike (and thus the onset of the resulting EPSP) precedes the postsynaptic spike, the shape of the theoretically derived learning window for physiologically plausible values of τ and ν_{max} ($\tau = 1 / \nu_{max} = 40$ ms; middle row in Figure 3) predicts potentiation of the synapse when a postsynaptic spike is preceded by the onset of an EPSP and depression of the synapse when this temporal order is reversed. This behavior is in agreement with experimental data from neocortex and hippocampus in rats as well as from the optic tectum in *Xenopus* [14–18]. To further illustrate this agreement, Figure 4 compares the data as published by Bi and Poo [16] with the learning window resulting from a smoothened power spectrum with the shape of a Cauchy function (Equation 35) instead of $\mathcal{P}_{f_{SFA}}$. As demonstrated above, this corresponds to implementing the slowness principle in form of the trace rule. Interestingly, the resulting learning window has the double-exponential shape that is regularly used in models of STDP (e.g., [24,29,30]). As the absolute scale of the learning window is not determined in our analysis, it was adjusted to facilitate the comparison with the experimental data.

Interpretation of the learning windows. The last section leaves a central question open: why are these learning windows optimal for slowness learning and why does the EPSP play such an important role for the shape of the learning window?

Let us first discuss the case of the symmetric learning window, that is, the situation in which the EPSP is shorter than the fastest timescale in the input signal. Then, the convolution with the EPSP has practically no effect on the temporal structure of the signal and the output firing rate can be regarded as an instantaneous function of the input rates. We can thus neglect the EPSP altogether. The learning mechanism can then be understood as follows: assume at a given time t the postsynaptic firing rate r^{out} is high and causes a postsynaptic spike. Then, the finite width of the learning window leads to potentiation not only of those synapses that participated in initiating the spike but also of those that transmit a spike within a certain time window around the time of the postsynaptic spike. As this leads to an increase of the firing rate within this time window, the learning mechanism tends to equilibrate the firing rates for neighboring times and thus favors temporally slow output signals.
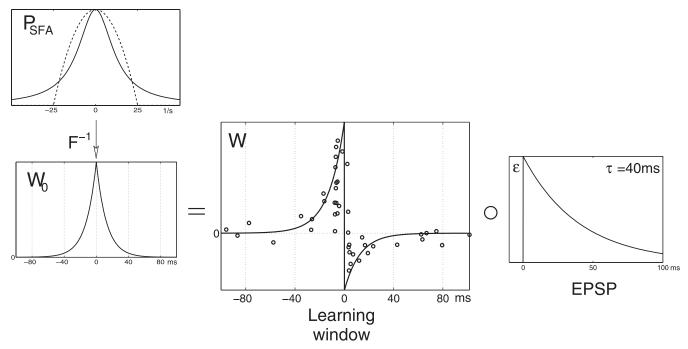
Figure 4. Comparison of the Learning Window with Experimental Data

The plot compares the theoretically predicted learning window with experimental data from hippocampal pyramidal cells as published by Bi and Poo [16] (larger plot in the middle). Instead of the ideal power spectrum $\mathcal{P}_{f_{SFA}}$ with the abrupt cutoff at v_{max} as stated in Equation 20, a Cauchy function with $\gamma = 1 / 15$ ms was used (top left, the dashed line is $\mathcal{P}_{f_{SFA}}$ for $v_{max} = 1 / (40$ ms)). Again, the EPSP decay time was $\tau = 40$ ms. This learning window corresponds to an implementation of the "trace rule" [1,4,6] for a decay time of the exponential filter of 15 ms.

doi:10.1371/journal.pcbi.0030112.g004

If the duration of the EPSP is longer than the fastest timescale in the input signal, the output firing rate is no longer an instantaneous function of the input signals but generated by low-pass filtering the signal a^{out} with the EPSP. This affects learning, because the objective of the continuous model is to optimize the slowness of a^{out}, whose temporal structure is now "obscured" by the EPSP. In order to optimize the objective, the system thus has to develop a deconvolution mechanism to reconstruct a^{out}. From this point of view, the learning window has to perform two tasks simultaneously. It has to first perform the deconvolution and then enforce slowness on the resulting signal. This is most easily illustrated by means of the condition in Equation 55. The convolution of the learning window with the EPSP generates the effective learning window W_0 that is independent of the EPSP and which coincides with the learning window for infinitely short EPSPs. Intuitively, we could solve Equation 55 by choosing a learning window that consists of the "inverse" of the EPSP and the EPSP-free learning window W_0. An intuitive example is the limiting case of an infinitely long EPSP. The EPSP then corresponds to a Heaviside function and performs an integration, which can be inverted by taking the derivative. Thus, the learning window for long EPSPs is the temporal derivative of the learning window for short EPSPs. The dependence of the required learning window on the shape of the EPSP is thus caused by the need of the learning window to "invert" the EPSP.

These considerations shed a different light on the shape of physiologically measured learning windows. The antisymmetry of the learning window may not act as a physiological implementation of a causality detector after all, but rather as

a mechanism for compensating intrinsic low-pass filters in neuronal processing such as the EPSP. For functional interpretations of STDP, it may be more sensible to consider the convolution of the learning window with the EPSP than the learning window alone.

It should be noted that, according to our learning rule, the weights adapt in order to make a hypothetical instantaneous output signal a^{out} optimally slow. This does not necessarily imply that the output firing rate r^{out}, which is generated by low-pass filtering a^{out} with the EPSP, is optimally slow. In principle, the system could generate more slowly varying signals by exploiting the temporal structure of the EPSP. However, the motivation for the slowness principle is the idea that the system learns to detect invariances in the *input* signal, and that from this perspective the goal of creating a slowly varying output signal is not an end in itself but a means to learn invariances. Thus, the low-pass filtering effect of the EPSP should not be exploited but ignored or compensated.

General learning windows and EPSPs. Although the asymmetry in LTP/LTD induction observed by Bi and Poo [16] has also been observed in other studies, the decay times for the LTP and the LTD branches of the learning window appear to be different in other preparations [18]. One may thus ask how robust our interpretation is with respect to the detailed shape of the learning window. To address this question, we start with some general learning window W and EPSP ε and ask under which conditions the effective learning window $W_0 = W \circ \varepsilon$ prefers slowly varying features in the input.

As a starting point, we use the dynamics of the weights in Equation 54 as generated by the input statistics. Using

$a^{\text{out}} = \sum_j w_j a_j^{\text{in}}$ and defining the correlation functions $C_{ij}(t) = [a_j^{\text{in}} * a_i^{\text{in}}](t)$ yields

$$\langle \Delta w_i \rangle_{E^{\text{in}}, E^{\text{out}}} = \sum_j \underbrace{\left[\gamma \int W_0(t) C_{ij}(t)\, dt \right]}_{=:A_{ij}} w_j. \tag{59}$$

The dynamics thus follows a linear difference equation with a dynamic matrix A_{ij} whose properties are determined by the correlation function $C_{ij}(t)$ and the effective learning window $W_0(t)$. One important question is whether the weights approach a stable fixed-point state or oscillate. In this context, the symmetry properties of A_{ij} and thus those of C_{ij} are crucial. The correlation functions obey the relation

$$C_{ij}(t) = C_{ji}(-t), \tag{60}$$

which couples their spatial symmetry (i.e., the symmetry with respect to the indices i and j) to their temporal symmetry. For instance, if the input statistics are reversible, i.e., for $C_{ij}(t) = C_{ij}(-t)$, C_{ij} is symmetric in the indices and so is A_{ij}. If the input statistics were "perfectly irreversible," i.e., $C_{ij}(t) = -C_{ij}(-t)$, C_{ij} and A_{ij} would be antisymmetric. This motivates the splitting of the correlation functions C_{ij} into a temporally symmetric and an antisymmetric component: $C_{ij} = C_{ij}^+ + C_{ij}^-$ with $C_{ij}^\pm(t) = \pm C_{ij}^\pm(-t)$. In a similar fashion, we split the effective learning window $W_0 = W_0^+ + W_0^-$. For symmetry reasons, the dynamical matrix A_{ij} can then be separated into two components

$$A_{ij} = \underbrace{\gamma \int W_0^+(t) C_{ij}^+(t)\, dt}_{=:A_{ij}^+} + \underbrace{\gamma \int W_0^-(t) C_{ij}^-(t)\, dt}_{=:A_{ij}^-}. \tag{61}$$

Because of the symmetry relation in Equation 60, A_{ij}^+ is symmetric in i and j, while A_{ij}^- is antisymmetric. This shows that the effective learning window W_0 can be split into two functionally different components. The symmetric component picks up the reversible aspects of the input statistics while the antisymmetric component detects irreversibilities, e.g., possible causal relations within the input data. It is this antisymmetric component of the learning window that has previously been interpreted as a means for sequence learning and predictive coding [19,31]. Note that the associated weight update $\sum_j A_{ij}^- w_j$ is always orthogonal to the weight itself. Thus, irreversibilities in the input data in combination with an antisymmetric learning window work against the development of a stable weight distribution, even if the input statistics are stationary. In particular, weight oscillations on the timescale of learning may occur. For instance, in networks with recurrent connections that learn according to STDP, previous studies have shown that the network tends to develop a state of distributed synchrony [32] that resembles synfire chains. These activity patterns display a pronounced causal structure, so it would be interesting to check if the synaptic weights that emerge in such a network are stable or show oscillations. It is likely that in this context the model constraints on the weights play an important role. If the weights are limited by hard boundaries as in [32], they tend to saturate, thereby avoiding oscillatory solutions. In the case of softer weight constraints, e.g., in models of STDP with multiplicative weight-dependence, oscillations may occur.

If W_0 is symmetric or if the input statistics are reversible,

$C_{ij}^- = 0$, the dynamical matrix $A_{ij} = A_{ij}^+$ is symmetric. As already seen for the case of the continuous model neuron, the learning dynamics can then be interpreted as a gradient ascent on the objective function

$$\Psi = \frac{1}{2} \sum_{i,j} w_i A_{ij}^+ w_j = \frac{1}{2} \int W_0^+(\nu) \mathcal{P}_{a^{\text{out}}}(\nu)\, d\nu. \tag{62}$$

As discussed earlier, this objective function can be interpreted as an implementation of the slowness principle if $W_0^+(\nu)$ is a low-pass filter, i.e., it has a global maximum at zero frequency. This indicates that at least for reversible input statistics the preference of STDP for slow signals may be rather insensitive to details of the learning window.

Discussion

Neurons in the central nervous system display a wide range of invariances in their response behavior, examples of which are phase invariance in complex cells in the early visual system [33], head direction invariance in hippocampal place cells [34], or more complex invariances in neurons associated with face recognition [35]. If these invariances are learned, the associated learning rule must somehow reflect a heuristics as to which sensory stimuli are supposed to be categorized as being the same. Objects in our environment are unlikely to change completely from one moment to the next but rather undergo typical transformations. Intuitively, responses of neurons with invariances to these transformations should thus vary more slowly than others. The slowness principle uses this intuition and conjectures that neurons learn these invariances by favoring slowly varying output signals without exploiting low-pass filtering.

SFA [10] is one implementation of the slowness principle in that it minimizes the mean square of the temporal derivative of the output signal for a given set of training data. SFA has been used to model a wide range of physiologically observed properties of complex cells in primary visual cortex [8] as well as translation, rotation, and other invariances in the visual system [10]. In combination with a sparse coding objective, SFA has also been used to describe the self-organized formation of place cells in the hippocampal formation [11].

The algorithm that underlies SFA is rather technical, and it has not yet been examined whether it is feasible to implement SFA within the limitations of neuronal circuitry. In this paper we approach this question analytically and demonstrate that such an implementation is possible in both continuous and spiking model neurons.

In the first part of the paper, we show that for linear continuous model neurons, the slowest direction in the input signal can be learned by means of Hebbian learning on low-pass filtered versions of the input and the output signal. The power spectrum of the low-pass filter required for implementing SFA can be derived from the learning objective and has the shape of an upside-down parabola.

The idea of using low-pass filtered signals for invariance learning is a feature that our model has in common with several others [1,4,6]. By means of the continuous model neuron, we have discussed the relation of our model to these "trace rules" and have shown that they bear strong similarities.

The second part of the paper discusses the modifications

that have to be made to adjust the learning rule for a Poisson neuron. We find that in an ensemble-averaged sense it is possible to reproduce the behavior of the continuous model neuron by means of spike-timing–dependent plasticity (STDP). Our study suggests that the outcome of STDP learning is not governed by the learning window alone but rather by the convolution of the learning window with the EPSP, which is of relevance for functional interpretations of STDP.

The learning window that realizes SFA can be calculated analytically. Its shape is determined by the interplay of the duration of the EPSP and the maximal input frequency ν_{max}, above which the input signals are assumed to have negligible power. If ν_{max} is small, i.e., if the EPSP is sufficiently short to temporally resolve the most quickly varying components of the input data, the learning window is symmetric, whereas for large ν_{max} or long EPSPs, it is antisymmetric. Interestingly, physiologically plausible parameters lead to a learning window whose shape and width is in agreement with experimental findings. Based on this result, we propose a new functional interpretation of the STDP learning window as an implementation of the slowness principle that compensates for neuronal low-pass filters such as the EPSP.

An important question in this context is on which timescales is this interpretation valid. It is conceivable that for signals that vary on a timescale of less than a hundred milliseconds, a learning window with a width of tens of milliseconds can distinguish slower from faster signals. STDP could thus be sufficient to establish invariant representations in early sensory processing, e.g., visual receptive fields that become invariant to microsaccades inducing small translations. Although it is unlikely that STDP alone can distinguish between signals that vary on behavioral timescales of hundreds of milliseconds or even seconds, this may not be problematic, because it is probably not sensible to order *all* aspects of the stimuli according to how quickly they vary. Rather, one should distinguish input components that vary so quickly that they are unlikely to be behaviorally relevant from those that vary on behavioral timescales. From this perspective, the intrinsic timescale of the learning rule should be such that its discriminative power is best on a timescale where this transition occurs. It is conceivable that this transition timescale lies on the order of several tens of milliseconds. The learning of high level invariances that correspond to behavioral timescales will probably require additional mechanisms with corresponding intrinsic timescales, e.g., sustained firing in response to a stimulus [36].

For general learning windows and EPSPs, the convolution of the learning window with the EPSP can be split into a symmetric component and an antisymmetric component. The symmetric component picks up reversible aspects of the input statistics while the antisymmetric component detects irreversible aspects. Previous functional interpretations of STDP have mostly concentrated on the antisymmetric component, which has been interpreted, e.g., as a mechanism for sequence learning or predictive coding [19,31] or for reducing recurrent connectivity in favor of feed-forward structures [30,32]. Other studies have neglected the phase structure of the learning window altogether and concentrated on its power spectrum, proposing that timing-dependent plasticity performs Hebbian learning on an optimal estimate of the input signals in the presence of noise [37,38].

Note that these interpretations are not necessarily contradictory to ours, because the slowness interpretation relies on the symmetric component of the learning window only and thus on the reversible aspect of the input statistics. These considerations indicate that depending on the temporal structure of the input, STDP may have different functional roles.

A different approach to unsupervised learning of invariances with a biologically realistic model neuron has been taken by Körding and König [39]. In their model, bursts of backpropagating spikes gate synaptic plasticity by providing sufficient amounts of dendritic depolarization. These bursts are assumed to be triggered by lateral connections that evoke calcium spikes in the apical dendrites of cortical pyramidal cells.

Of course the model presented here is not a complete implementation of SFA. We have only considered the central step of SFA, the extraction of the most slowly varying direction from a set of whitened input signals. To implement the full algorithm, additional steps are necessary: a nonlinear expansion of the input space, the whitening of the expanded input signals, and a means of normalizing the weights. When traversing the dendritic arborizations of a postsynaptic neuron, axons often make more than one synaptic contact. As different input channels may be subjected to different nonlinearities in the dendritic tree (cf. [40]), the postsynaptic neuron may have access to several nonlinearly transformed versions of the same presynaptic signals. Conceptually, this resembles a nonlinear expansion of the input signals. However, it is not obvious how these signals could be whitened within the dendrite. On the network level, however, whitening could be achieved by adaptive recurrent inhibition between the neurons [41]. This mechanism may also be suitable for extracting several slow uncorrelated signals as required in the original formulation of SFA [10] instead of just one. We assumed an explicit weight normalization in the description of our model. However, one could also use a modified learning rule that implicitly normalizes the weight vector as long as it extracts the signal with the largest variance. A possible biological mechanism is synaptic scaling [25], which is believed to multiplicatively rescale all synaptic weights according to postsynaptic activity, similar to Oja's rule [26,42]. Thus, it appears that most of the mechanisms necessary for an implementation of the full SFA algorithm are available, but that it is not yet clear how to combine them in a biologically plausible way.

Another critical point in the analytical derivation for the spiking model is the replacement of the temporal by the ensemble average, as this allows recovery of the rates that underlie the Poisson processes. The validity of the analytical results thus requires some kind of ergodicity in the training data, a condition which of course needs to be justified for the specific input data at hand.

It is still open whether the results presented here can be reproduced with more realistic model neurons. The spiking model neuron used here was simplified in that it had a linear relationship between input and output firing rate. In many real neurons, highly nonlinear behavior was observed. Interestingly, Hebbian learning for nonlinear rate-based neurons has previously been associated with the detection of higher-order moments of the input statistics [43], thereby providing a mechanism for extracting statistically independ-

ent components of the input signal. Because for sparse input statistics independent component analysis is closely related to sparse coding [44], it is tempting to speculate that within a rate picture, temporally nonlocal plasticity with a nonlinear input–output relation implements a combination of sparseness and slowness. Learning paradigms that combine these two objectives are thus an interesting field for further studies [11,45].

Another nonlinearity that we have neglected is the frequency- and weight-dependence of STDP [16,46]. Additional work will be needed to examine how these interfere with the proposed functional role of STDP. Furthermore, modeling the spiking mechanism of a neuron by an inhomogeneous Poisson process is also a severe simplification that ignores basic phenomena of spike generation in biological neurons such as refractoriness and thresholding. It is not clear how these characteristics would change the learning rule that leads to an implementation of the slowness principle. It seems to be a very difficult task to answer these questions analytically. Simulations will be necessary to verify the results derived here and to analyze which changes appear and which adaptations must be made in a more realistic model of neural information processing.

In summary, the analytical considerations presented here show that (i) slowness can be equivalently achieved by minimizing the variance of the time derivative signal or by maximizing the variance of the low-pass filtered signal, the latter of which can be achieved by standard Hebbian learning on the low-pass filtered input and output signals; (ii) the difference between SFA and the trace learning rule lies in the exact shape of the effective low-pass filter—for most practical purposes the results are probably equivalent; (iii) for a spiking Poisson model neuron with an STDP learning rule, it is not the learning window that governs the weight dynamics but the convolution of the learning window with the EPSP; (iv) the STDP learning window that implements the slowness objective is in good agreement with learning windows found experimentally. With these results, we have reduced the gap between slowness as an abstract learning principle and biologically plausible STDP learning rules, and we offer a completely new interpretation of the standard STDP learning window.

Methods

The methods employed in this paper rely on standard mathematical techniques as commonly used in the theory of synaptic plasticity (see, e.g., [47]).

Acknowledgments

References

1. Földiak P (1991) Learning invariance from transformation sequences. Neural Comput 3: 194–200.
2. Mitchison G (1991) Removing time variation with the anti-Hebbian differential synapse. Neural Comput 3: 312–320.
3. Becker S, Hinton GE (1992) A self-organizing neural network that discovers surfaces in random-dot stereograms. Nature 355: 161–163.
4. O'Reilly RC, Johnson MH (1994) Object recognition and sensitive periods: A computational analysis of visual imprinting. Neural Comput 6: 357–389.
5. Stone JV, Bray A (1995) A learning rule for extracting spatio–temporal invariances. Network: Comput Neural Sys 6: 429–436.
6. Wallis G, Rolls ET (1997) Invariant face and object recognition in the visual system. Prog Neurobiol 51: 167–194.
7. Peng HC, Sha LF, Gan Q, Wei Y (1998) Energy function for learning invariance in multilayer perceptron. Electronics Lett 34: 292–294.
8. Berkes P, Wiskott L (2005) Slow feature analysis yields a rich repertoire of complex cells. J Vis 5: 579–602.
9. Körding KP, Kayser C, Einhäuser W, König P (2004) How are complex cell properties adapted to the statistics of natural stimuli? J Neurophysiol 91: 206–212.
10. Wiskott L, Sejnowski T (2002) Slow feature analysis: Unsupervised learning of invariances. Neural Comput 14: 715–770.
11. Franzius M, Sprekeler H, Wiskott L (2007) Unsupervised learning of place cells, head direction cells, and spatial-view cells with slow feature analysis on quasi-natural videos. Cognitive Sciences EPrint Archive (CogPrints) 5492. Available: http://cogprints.org/5492/. Accessed 4 June 2007.
12. Wyss R, König P, Verschure PFMJ (2006) A model of the ventral visual system based on temporal stability and local memory. PLoS Biol 4: e120.
13. Kempter R, Gerstner W, van Hemmen JL (1999) Hebbian learning and spiking neurons. Phys Rev E 59: 4498–4514.
14. Debanne D, Gähwiler BH, Thomson SM (1994) Asynchronous pre- and postsynaptic activity induces associative long-term depression in area CA1 of the rat hippocampus. Proc Natl Acad Sci U S A 91: 1148–1152.
15. Markram H, Lübke J, Frotscher M, Sakmann B (1997) Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. Science 275: 213–215.
16. Bi G-q, Poo M-m (1998) Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. J Neurosci 18: 10464–10472.
17. Zhang LI, Tao HW, Holt CE, Harris WA, Poo M-m (1998) A critical window for cooperation and competition among developing retinotectal synapses. Nature 395: 37–44.
18. Feldman DE (2000) Timing-based LTP and LTD at vertical input to layer II/III pyramidal cells in rat barrel cortex. Neuron 27: 45–56.
19. Abbott LF, Blum KI (1996) Functional significance of long-term potentiation for sequence learning and prediction. Cereb Cortex 6: 406–416.
20. Gerstner W, Kempter R, van Hemmen JL, Wagner H (1996) A neuronal learning rule for sub-millisecond temporal coding. Nature 383: 76–78.
21. Kepecs A, van Rossum MCW, Song S, Tegner J (2002) Spike-timing–dependent plasticity: Common themes and divergent vistas. Biol Cybern 87: 446–458.
22. Kistler WM, van Hemmen JL (2000) Modeling synaptic plasticity in conjunction with the timing of pre- and postsynaptic action potentials. Neural Comput 12: 385.
23. Rubin J, Lee DD, Sompolinsky H (2001) Equilibrium properties of temporally asymmetric Hebbian learning. Phys Rev Lett 86: 364–367.
24. Gütig R, Aharonov S, Rotter S, Sompolinsky H (2003) Learning input correlations through nonlinear temporally asymmetric Hebbian plasticity. J Neurosci 23: 3697–3714.
25. Turrigiano GG, Nelson SB (2000) Hebb and homeostasis in neuronal plasticity. Curr Opin Neurobiol 10: 358–364.
26. Abbott LF, Nelson SB (2000) Synaptic plasticity: Taming the beast. Nat Neurosci 3: 1178–1183.
27. Kempter R, Gerstner W, van Hemmen JL (2001) Intrinsic stabilization of output rates by spike-based Hebbian learning. Neural Comput 13: 2709–2741.
28. Koch C, Rapp M, Segev I (1996) A brief history of time (constants). Cereb Cortex 6: 92–101.
29. van Rossum MCW, Bi G-q, Turrigiano GG (2000) Stable Hebbian learning from spike-timing–dependent plasticity. J Neurosci 20: 8812–8821.
30. Song S, Abbott LF (2001) Cortical mapping and development through spike-timing–dependent plasticity. Neuron 32: 339–350.
31. Rao RPN, Sejnowski TJ (2001) Spike-timing–dependent Hebbian plasticity as temporal difference learning. Neural Comput 13: 2221–2238.
32. Horn D, Levy N, Meilijson I, Ruppin E (2000) Distributed synchrony of spiking neurons in a Hebbian cell assembly. In: Müller K-R, editor. Adv Neural Info Process Syst (NIPS) 12. Cambridge (Massachusetts): MIT Press.
33. Hubel D, Wiesel T (1968) Receptive fields and functional architecture of monkey striate cortex. J Physiol (Lond) 195: 215–243.
34. Muller R, Bostock E, Taube JS, Kubie JL (1994) On the directional firing properties of hippocampal place cells. J Neurosci 14: 7235–7251.

35. Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I (2005) Invariant visual representation by single neurons in the human brain. Nature 435: 1102–1107.

36. Drew PJ, Abbott LF (2006) Extending the effects of spike-timing–dependent plasticity to behavioral timescales. Proc Natl Acad Sci U S A 103: 8876–8881.

37. Wallis G, Baddeley R (1997) Optimal, unsupervised learning in invariant object recognition. Neural Comput 9: 883–894.

38. Dayan P, Häusser M, London M (2004) Plasticity kernels and temporal statistics. In: Schölkopf B, editor. Adv Neural Info Process Syst (NIPS) 16. Cambridge (Massachusetts): MIT Press.

39. Körding KP, König P (2001) Neurons with two sites of synaptic integration learn invariant representations. Neural Comput 13: 2823–2849.

40. London M, Häusser M (2005) Dendritic computation. Ann Rev Neurosci 28: 503–532.

41. Barlow H, Földiak P (1989) Adaptation and decorrelation in the cortex. In: Durbin R, Miall C, Mitchison G. Computing neuron. New York: Addison-Wesley. 43 p.

42. Oja E (1982) A simplified neuron as a principal component analyzer. J Math Biol 15: 267–273.

43. Oja E, Karhunen J (1995) Signal separation by nonlinear Hebbian learning. Computational intelligence: A dynamic system perspective: 83–97.

44. Olshausen BA, Field DJ (1997) Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision Res 37: 3311–3325.

45. Blaschke T, Zito T, Wiskott L (2007) Independent slow feature analysis and nonlinear blind source separation. Neural Comput 19: 994–1021.

46. Sjöström PJ, Turrigiano GG, Nelson SB (2001) Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. Neuron 32: 1149–1164.

47. Gerstner W, Kistler WM (2002) Spiking model neurons. Cambridge University Press.